RAMP: collaborative data challenges run by the Paris-Saclay Center for Data Science

C. Marini[1], Balazs Kegl[1], Alexandre Gramfort[2]

[1]CNRS, [2]Institut Mines Telecom

Paris-Saclay Center for Data Science

http://www.datascience-paris-saclay.fr/

250 researchers in **35** laboratories

Biology & bioinformatics
IBISC/UEvry
LRI/UPSud
Hepatinov
CESP/UPSud-UVSQ-Inserm
IGM-I2BC/UPSud
MIA/Agro
MIAj-MIG/INRA
LMAS/Centrale

Chemistry
EA4041/UPSud

Earth sciences
LATMOS/UVSQ
GEOPS/UPSud
IPSL/UVSQ
LSCE/UVSQ
LMD/Polytechnique

Economy
LM/ENSAE
RITM/UPSud
LFA/ENSAE

Neuroscience
UNICOG/Inserm
U1000/Inserm
NeuroSpin/CEA

Particle physics astrophysics & cosmology
LPP/Polytechnique
DMPH/ONERA
CosmoStat/CEA
IAS/UPSud
AIM/CEA
LAL/UPSud

Machine learning
LRI/UPSud
LTCI/Telecom
CMLA/Cachan
LS/ENSAE
LIX/Polytechnique
MIA/Agro
CMA/Polytechnique
LSS/Supélec
CVN/Centrale
LMAS/Centrale
DTIM/ONERA
IBISC/UEvry
LIST/CEA

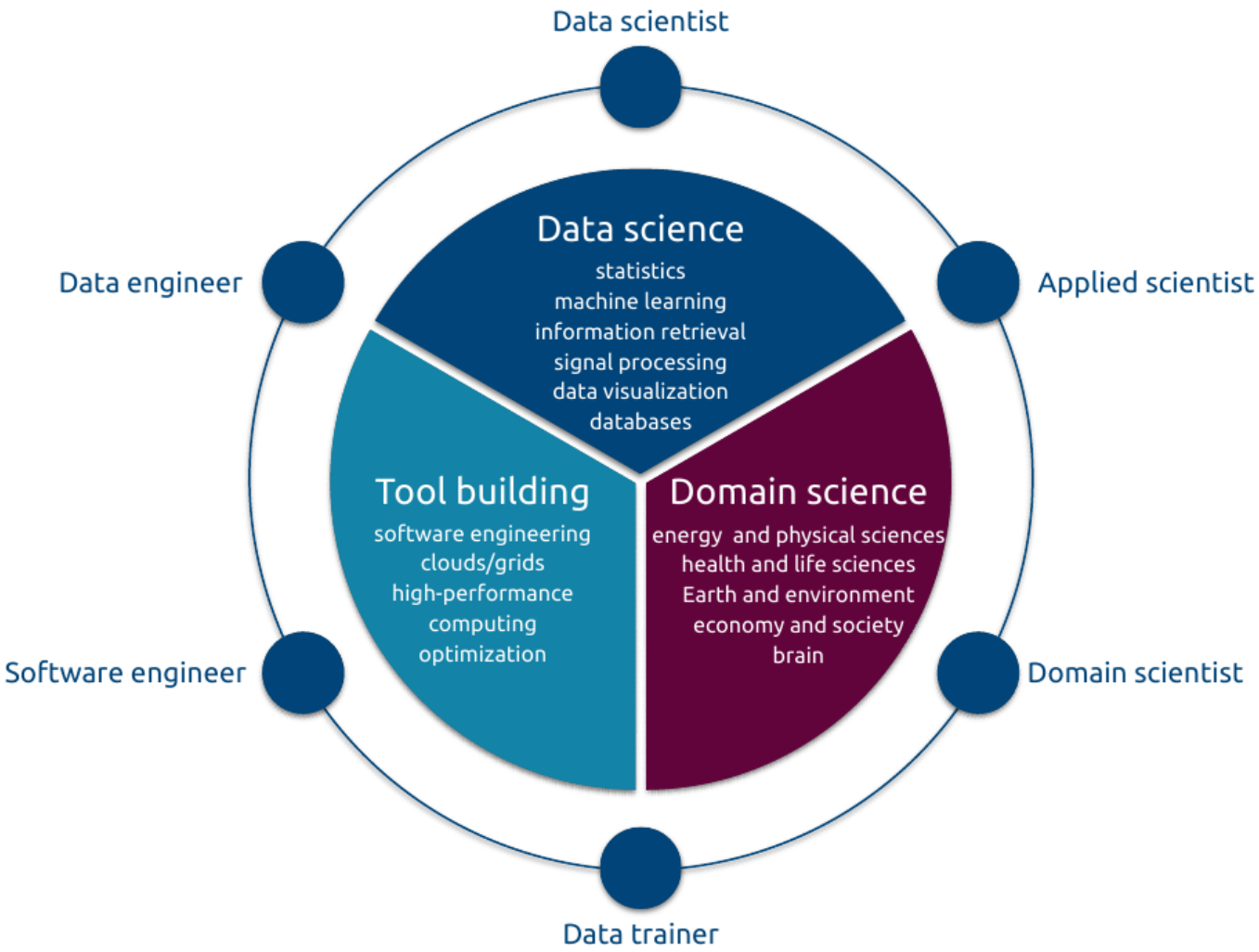
Visualization
INRIA
LIMSI

Signal processing
LTCI/Telecom
CMA/Polytechnique
CVN/Centrale
LSS/Supélec
CMLA/Cachan
LIMSI
DTIM/ONERA

Statistics
LMO/UPSud
LS/ENSAE
LSS/Supélec
CMA/Polytechnique
LMAS/Centrale
MIA/AgroParisTech

Paris-Saclay Center for Data Science

A multi-disciplinary initiative to define, structure, and manage the data science ecosystem at the University Paris-Saclay



RAMP Rapid Analytics and Model Prototyping

Collaborative Data Challenge (RAMP)

- Connection between data domain science and data science experts
- Training tool

RAMP Rapid Analytics and Model Prototyping

Collaborative Data Challenge (RAMP)

- Connection between data domain science and data science experts
- Training tool

RAMP lifecycle

Preparation:

- Domain expert brings: a **prediction problem** and associated **dataset**
- Data scientist helps: **formulate a machine learning problem** and **clean the data**

Event:

- Participants submit models (code) and can look at each other submission
- Models are trained on our backend
- Scores are on a leaderboard

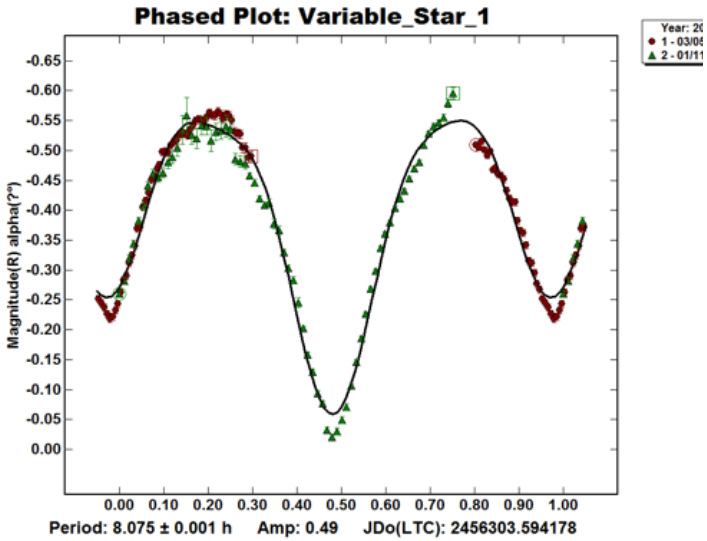
Follow-Up:

- Collaborative Paper, application

RAPID ANALYTICS AND MODEL PROTOTYPING

2015 Apr 10

Classifying variable stars



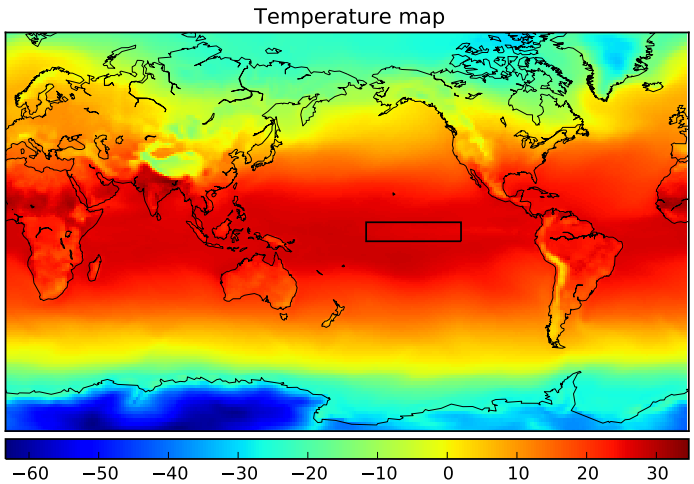
université PARIS-SACLAY Paris-Saclay Center for Data Science

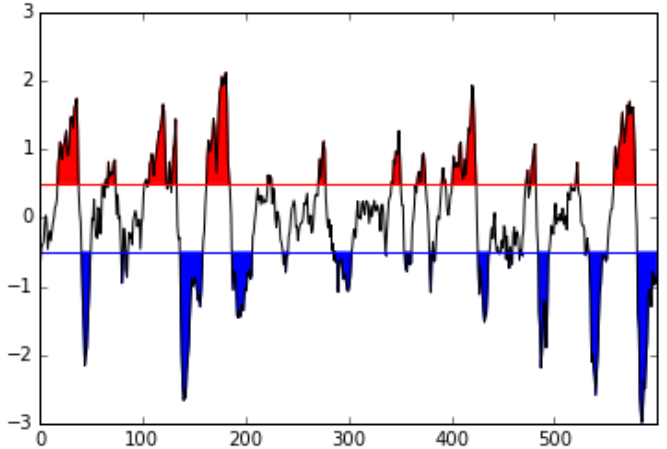
Accuracy improvement: 89 to 96%

RAPID ANALYTICS AND MODEL PROTOTYPING

2015 June 16 and Sept 26

Predicting El Nino





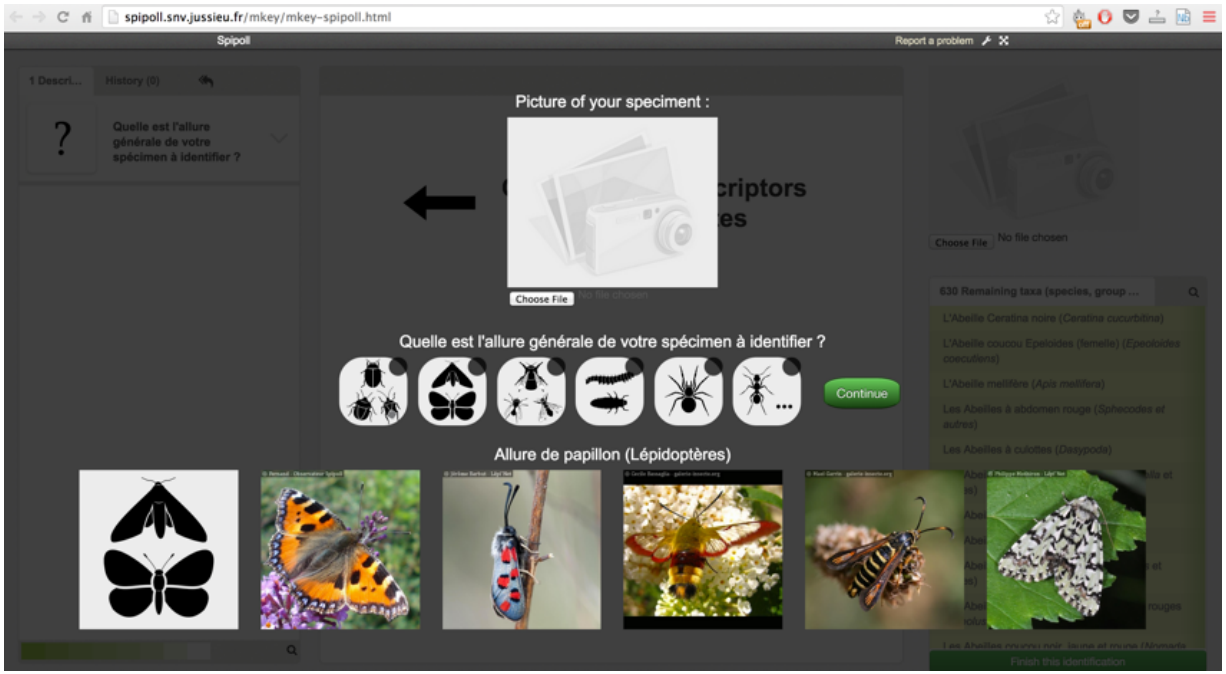
université PARIS-SACLAY Paris-Saclay Center for Data Science

RMSE improvement: 0.9°C to 0.4°C

RAPID ANALYTICS AND MODEL PROTOTYPING

2015 October 8

Insect classification



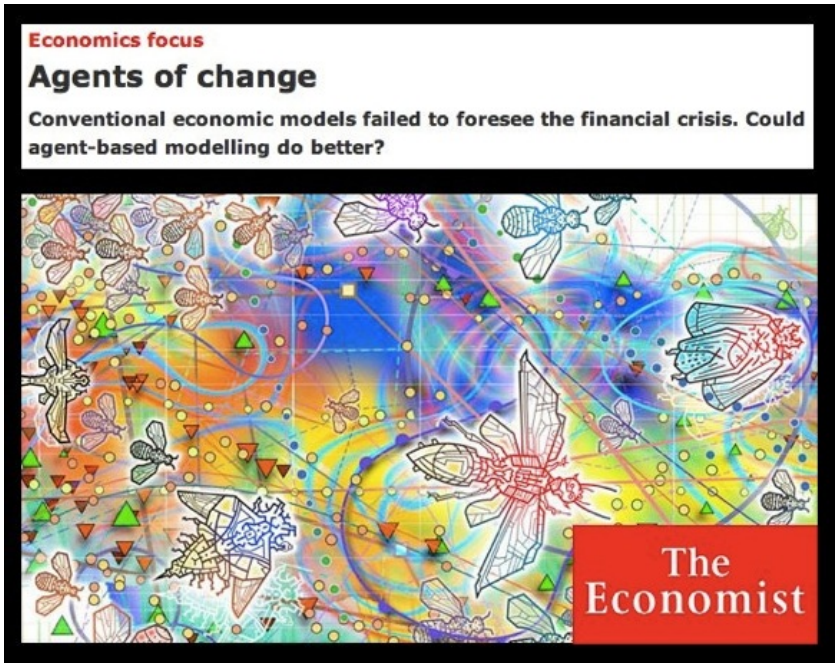
38

Accuracy improvement: 30 to 70%

RAPID ANALYTICS AND MODEL PROTOTYPING

2016 February 10

Macroeconomic agent-based models



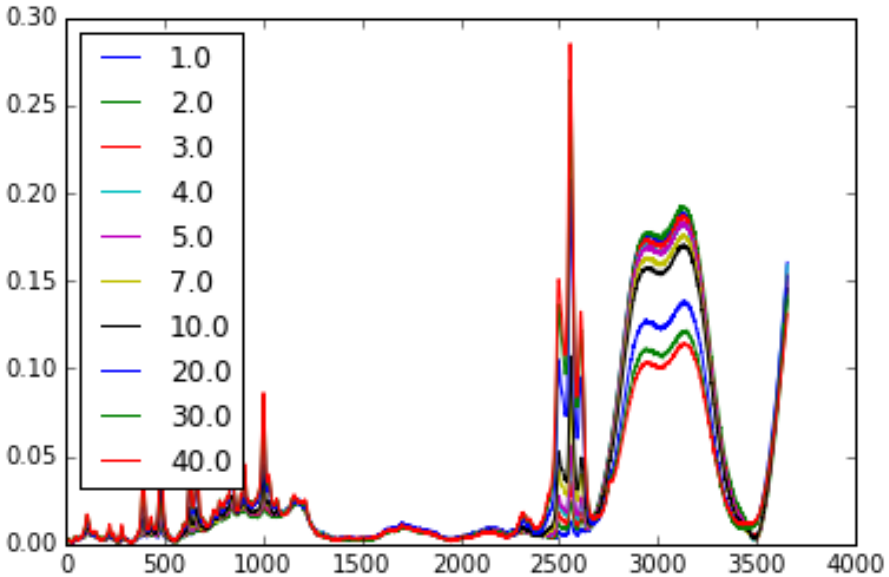
universite PARIS-SACLAY Paris-Saclay Center for Data Science

f1 score improvement: 0.57 to 0.63

RAPID ANALYTICS AND MODEL PROTOTYPING

2016 May 11

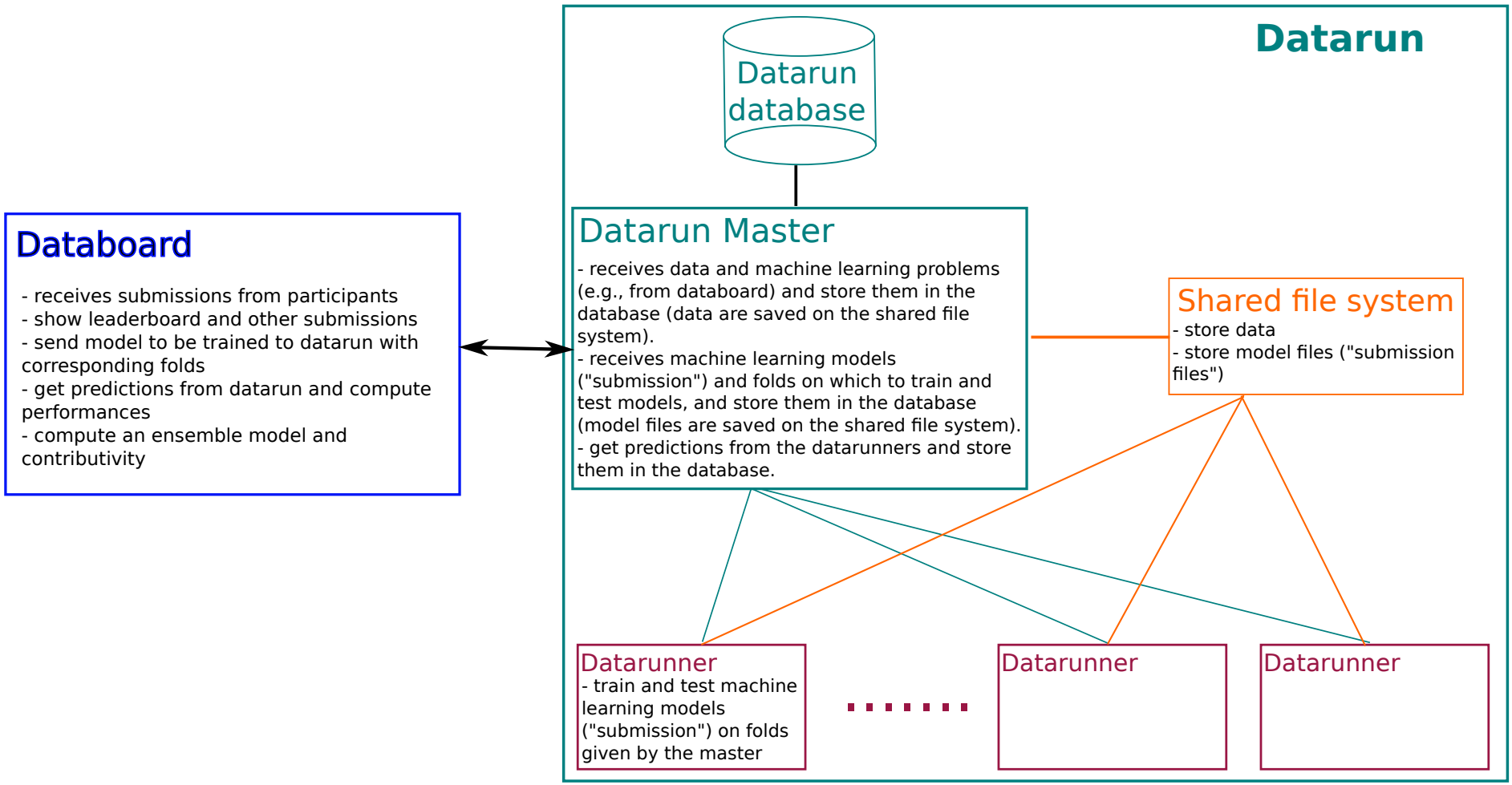
Drug identification from spectra



universite PARIS-SACLAY Paris-Saclay Center for Data Science

Drug detection accuracy improvement: 9 to 3%
Drug concentration mean abs rel error improvement: 20 to 12%

RAMP platforms



Databoard:

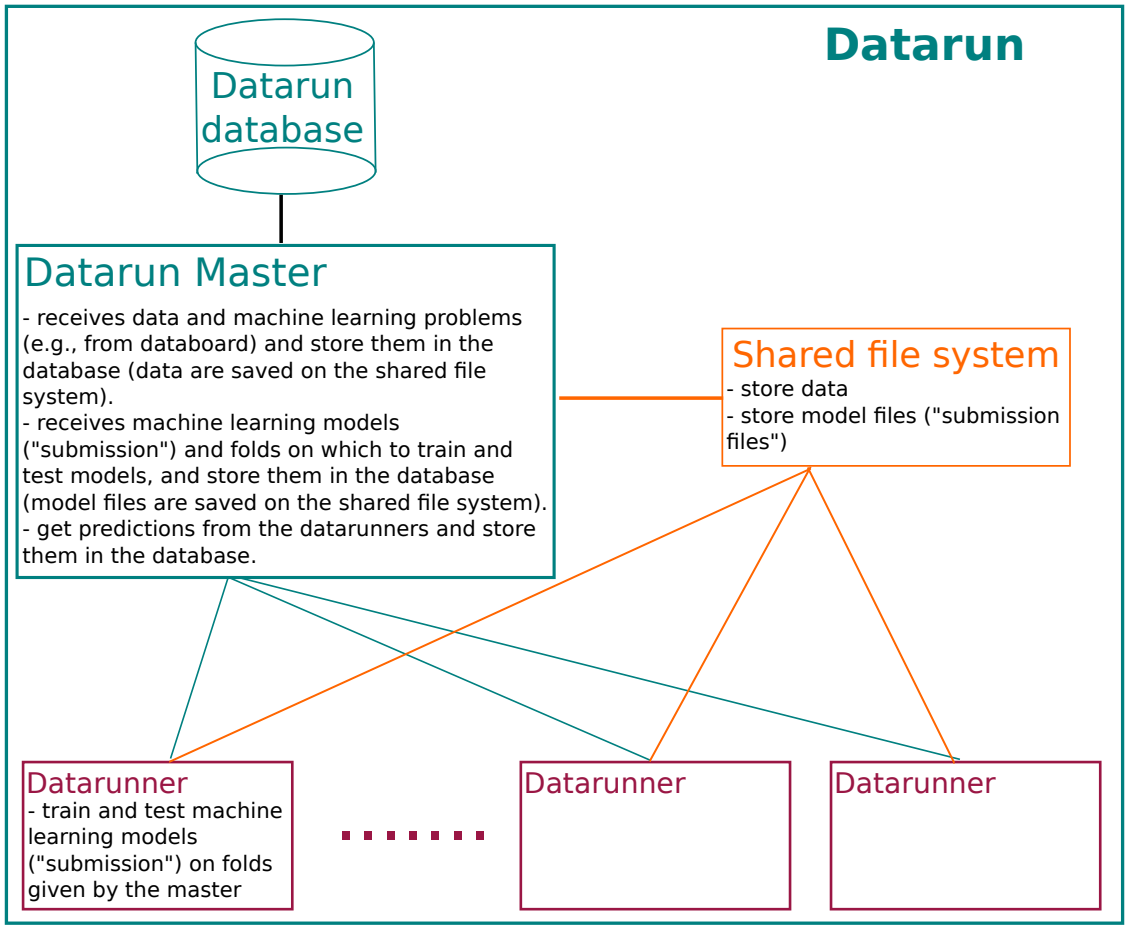
- Web application developped in Flask (Python)
- Deployed on VM from openstack (deployed at LAL)

Databoard

- receives submissions from participants
- show leaderboard and other submissions
- send model to be trained to datarun with corresponding folds
- get predictions from datarun and compute performances
- compute an ensemble model and contributivity

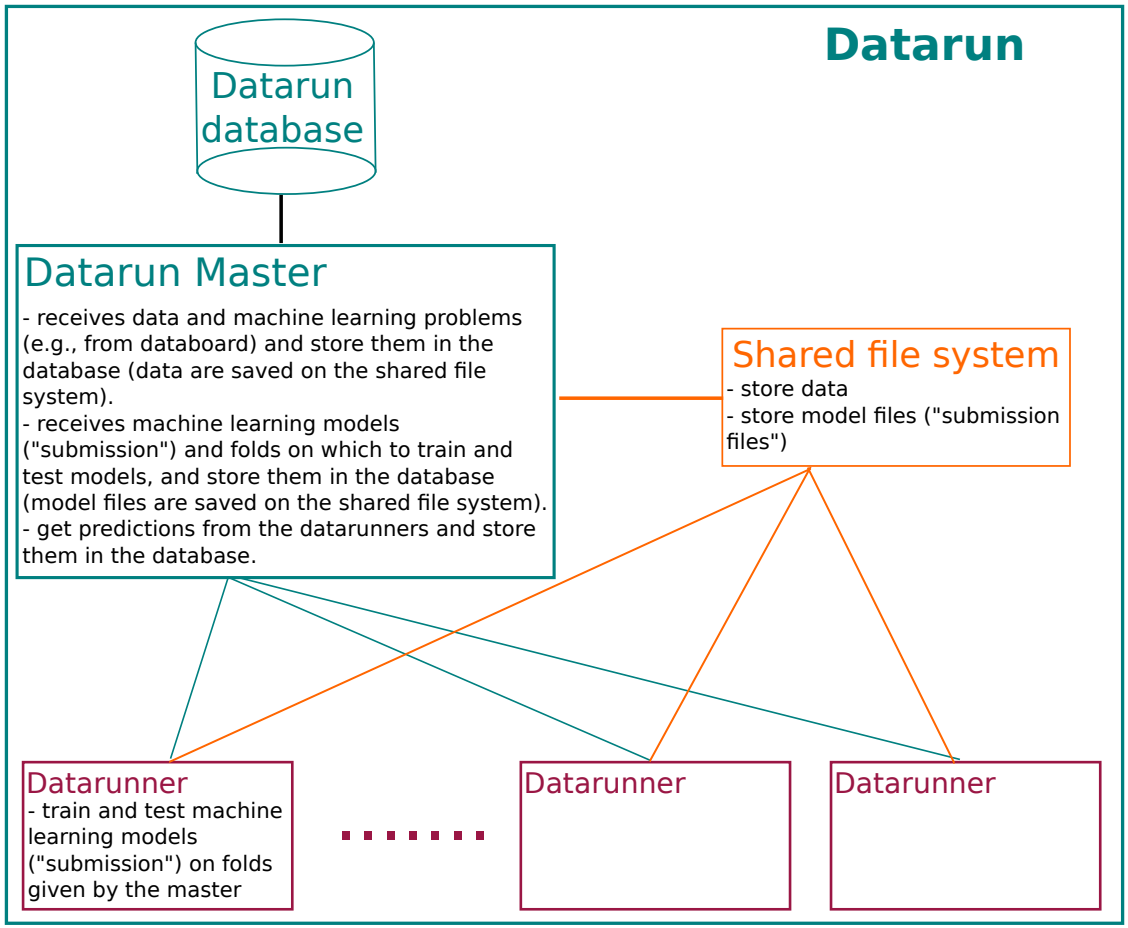
Datarun: train and test of machine learning models

- Web application developped in Django REST Framework (API REST) + Celery (datarunner management)
- Deployed on VMs from openstack (deployed at LAL)



Datarun: train and test of machine learning models

- Possible to start and stop datarunners to optimize computational resources
- Train and test from models written in Python



Demo

ramp.studio

Next steps

- Datarunners on GPU
- Use of container for datarunners
- Train and test of models written in languages different from Python
- Interface on databoard to easily set up a RAMP

Thanks!